Evaluating Criteria for English Learner Reclassification: A Causal-Effects Approach Using a Binding-Score Regression Discontinuity Design With Instrumental Variables

Joseph P. Robinson
University of Illinois at Urbana-Champaign

When English learners are “reclassified” as fluent English proficient, often their instructional setting changes, including a significant reduction in or elimination of English language development services. Depending on a child’s language skills, this instructional change could hinder future development or provide needed opportunities for learning advanced material. By establishing assessment-based guidelines for reclassification, policy makers have tremendous influence on when these settings change. The author highlights this policy lever for guiding reclassification decisions and identifies a method for rigorously evaluating whether the threshold for transitioning between settings is appropriate. This method—binding-score regression discontinuity with an instrumental variable—was then implemented to obtain unbiased effects of reclassification on academic outcomes for students on the cusp of meeting reclassification criteria to provide credible policy recommendations for maintaining or shifting assessment-based reclassification thresholds. The method detailed here can be used by policy makers to evaluate their own assessment-based guidelines.

Keywords: English language learner, reclassification, regression discontinuity design, policy evaluation

There are substantial and persistent Hispanic-White achievement gaps in mathematics and reading (see, e.g., Fryer & Levitt, 2004; Reardon & Robinson, 2008). Educators and policy makers need to be aware of the impact current educational policies have on these gaps. Hispanic students are a diverse group, with some fluent in English and others who are learning it. By definition, those in the latter group, English language learners (ELLs), are still acquiring basic communicative skills in and academic knowledge of English, which is a likely reason why the achievement gaps ELLs face are even greater than those faced by Hispanics as a group (August & Hakuta, 1997; Gándara, Rumberger, Maxwell-Jolly, & Callahan, 2003; Reardon & Robinson, 2008; Robinson, 2008).

In an effort to facilitate English proficiency, a common policy is for schools to provide auxiliary services or special coursework to ELLs. When an ELL attains fluency level, the child’s status is changed to reclassified fluent English proficient (R-FEP), and the instructional setting changes, including a reduction in or elimination of the aforementioned services and/or coursework (Gándara & Rumberger, 2006). Depending on a child’s level of English proficiency when the setting changes, reclassification may have different effects on future academic success. For example,

This research was funded through a grant from the Spencer Foundation. I thank Sean Reardon, Susanna Loeb, Kenji Hakuta, Andrei Cimpian, Amy Schwartz, Sarah Lubienski, and Daphna Bassok for helpful comments. Also, I thank the district for graciously sharing their data. All errors are my own.
if the instructional setting is changed when ELLs are still benefiting more from the ELL setting than they would in the R-FEP setting (e.g., if the English development services are still helpful), then we would observe that reclassified students perform more poorly on subsequent tests of English than their ELL counterparts. Conversely, if the English development services, a primary component of the ELL setting, were still provided to highly proficient ELLs, the time spent on these specialized linguistic support services might have been better apportioned to higher level academic content. Under these conditions, we might observe the new R-FEPs outperforming the ELLs. Thus, it is crucial that the policies designed to transition students between the different instructional settings ensure that the transition from ELL to R-FEP is smooth (e.g., no significant differences in performance). In this article, I describe a novel method that can be used by policy makers for evaluating the effects of reclassification and then illustrate how this method can be implemented, using data from a large urban school district in California.

Policy makers have tremendous influence on the reclassification process through the guidance they provide teachers via establishing assessment-based thresholds for reclassification (e.g., if an ELL scores above a certain threshold on an assessment, he or she becomes eligible for reclassification). Reclassification below the policy-specified assessment threshold is exceedingly rare. Once the threshold is attained, the probability of reclassification increases substantially, although teachers (and, sometimes, parents) do have some leeway in deciding whether children are reclassified. Given the importance of this threshold choice, policy makers can (and should) adopt guidelines to promote a smooth transition from ELL to R-FEP status (i.e., null effect), and thus, this study is focused on assessing the effects of the policy-specified assessment levels.

This raises an important distinction between this study and prior studies looking at reclassification effects: Other studies (e.g., Callahan, Wilkinson, & Muller, 2010; Flores, Painter, & Pachon, 2009) have reported average achievement differences between students receiving English language development (ELD) services and not receiving ELD services averaged over the range of ELD levels, while the current study is focused on the academic effects of reclassification at the assessment threshold (i.e., the policy lever). As discussed later in the “Methods” section, the approach of this study yields a stronger causal warrant than the approaches used in the existing literature, which are susceptible to selection bias on unobservable characteristics. Moreover, because it is not based on selection bias and deals with the direct policy mechanism for guiding reclassification decisions, the estimand of this study is more policy relevant.

I begin by arguing that reclassification is accompanied by a change in instructional setting (e.g., removal of specialized linguistic services) and, as such, should be viewed as a policy that could affect student outcomes. While reclassification could affect outcomes, I argue that if policy makers establish their assessment-based criteria at appropriate levels to facilitate reclassification decisions, then there should be no effect at the specified threshold, as both “positive” and “negative” effects suggest an inappropriate threshold for transition from the nonreclassified setting to the reclassified one. To evaluate the threshold, I propose a rigorous quasi-experimental method—binding-score regression discontinuity (RD) with instrumental variables (IVs)—that any state or district can use to evaluate its own assessment-based thresholds. After proposing this method, I implement it to evaluate the effects of one district’s specified thresholds on various student outcomes (current and subsequent English language arts achievement, attendance, and advanced course taking) and to illustrate how its estimates compare with those of other approaches, which condition on observable characteristics. I conclude with a discussion of how the methods can be extended to study cases of heterogeneity in reclassification effects.

**Background**

**Reclassification as Cause or Effect**

One of the primary goals of any program serving English learners is for ELLs to attain fluency in English. Reclassification, in its simplest form, is a label change from “ELL” to “R-FEP” when a student attains fluency level. Therefore, reclassifying all ELLs is in theory an admirable goal that we should be striving toward. Indeed, low reclassification rates before the passage of Proposition 227 in California were used to spur passage of the law that eliminated bilingual education as the default placement for new ELLs
(Grissom, 2004). And since its passage, researchers have used reclassification rates as outcome measures to evaluate Proposition 227 and similar laws (see, e.g., Grissom, 2004; Jepsen & de Alth, 2005; Thomas & Collier, 2005).

Rather than looking at reclassification as an outcome (i.e., reclassification as “effect”), in the current study, I examine the effects of reclassification on other outcomes (i.e., reclassification as “cause”). Although reclassification appears to be merely a label change, it actually is a change in instructional setting, including the elimination of or reduction in ELD services, thereby focusing these services on students with lower English proficiency, who presumably need them the most (Gándara & Rumberger, 2006; Linquanti, 2001).

**Reclassification Process**

California Education Code § 313(d) stipulates the use of multiple criteria in determining whether to reclassify a child from ELL to R-FEP status, but each district in California ultimately determines its own criteria for reclassification (cf., Gándara & Merino, 1993; Linquanti, 2001). The criteria suggested by the state include, but are not limited to, California English Language Development Test (CELDT) scores, teacher evaluations, parental opinion, and a comparison of basic skills relative to those of English-proficient students of the same age, interpreted by districts as standardized test scores on an assessment of English language arts (i.e., the California Standards Test [CST] of English Language Arts [CST-ELA]). Along with CST-ELA scores, some districts (but not the one studied here) also consider CST mathematics scores.

Schools receive the assessment-based criteria scores (i.e., CST and CELDT) by the academic midyear, and reclassification usually takes place near the start of the spring semester. Thus, reclassification occurs typically a few months prior to the next CST administration.

In addition to variation in criteria for reclassification, districts differ in implementation of instructional practices that change with reclassification (Parrish et al., 2006). All ELLs are required by California law to receive ELD instruction for a minimum of 30 minutes daily, but some districts provide more than 30 minutes to select ELLs, depending on their ELD levels.

**Services for ELLs in the District of Analysis**

The district used for illustration in this study established a set of criteria for services that should be provided to ELLs at different ELD levels. These are divided into two broader categories—ELD Levels 1 to 3 (lower English proficiency) and ELD Levels 4 and 5 (higher English proficiency)—with some differentiation among the lower group, providing more intensive services to ELD Levels 1 and 2. However, it is important to bear in mind that the question of this study (and its analytic method) concerns a causal-effect estimate for students close to meeting all test-based reclassification criteria, which is more often ELD Levels 4 and 5. A key difference between ELLs and R-FEPs is the amount of daily instruction on ELD, which is between 30 and 60 minutes (or 1 period) for ELD Levels 4 and 5 and 0 minutes for R-FEPs. That is, in this district, ELD instruction ceases for R-FEPs, who instead typically spend the 30 to 60 minutes on English language arts instruction activities.

Across all grade levels, teachers of ELLs must possess (or be in training for) a Crosscultural Language and Academic Development (CLAD) or Bilingual CLAD (BCLAD) certificate. This requirement affects the allocation of students across classrooms (i.e., ELLs are placed with CLAD or BCLAD teachers, while R-FEPs and other non-ELLs do not require teachers with these specialized credentials) and ELLs’ access to higher level courses (e.g., if a CLAD or BCLAD teacher does not teach advanced science courses, ELLs cannot enroll). In an effort to provide ELLs with access to college-preparatory courses, the district has created ELD versions of English language arts classes satisfying part of the English language arts (B, of A–G) California Regents Board requirement for entrance into a public university in California. This is in accordance with Regents policy allowing 1 year of ELD English to count toward the A–G requirements. No similar accommodations were made for other content areas; however, ELLs can enroll in college-preparatory courses if the teacher holds a CLAD or BCLAD certificate.

While the ELD instructional change is arguably the biggest service and condition difference between ELLs and R-FEPs, other conditions change as well (e.g., required teacher certification, classroom peer composition). This bundle
of service and condition changes cannot be disaggregated in this district, as they are all consequences of reclassification; thus, in this study, I estimate the treatment effect for the bundle of service and condition changes associated with the status change from ELL to R-FEP, not any one isolated component of the bundle or setting.

Previous Research on Reclassification and Related Topics

The literature that deals explicitly with reclassification as a whole is limited (but see Flores et al., 2009; Grissom, 2004; Jepsen & de Alth, 2005; Linquanti, 2001). For the purposes of this article, I focus on research describing the effects of specific aspects of instruction that change due to reclassification; these changes (in the district studied) were detailed in the previous subsection. The greatest change is likely to be the removal of daily ELD instruction, though this review also includes the existing literature on the broader differences between instructional settings for ELLs and R-FEPs and the expected consequences of those differences. To preview the conclusions, these studies are mixed, suggesting that ELD is potentially beneficial for elementary school students but that the track stratification for secondary school students can act as a barrier to advanced coursework. One important thing to bear in mind during this review is that the existing literature has focused on estimating the effects of ELD services across a spectrum of ELD levels, but in the current study, I estimate the effect at the point of the district’s reclassification threshold.

Regrettably, the literature on the effects of ELD instruction, or the effects of removing ELD instruction, is minimal. Recent reviews have attempted to determine which types of programs (e.g., ELD or English as a second language [ESL], bilingual education, mainstream English) result in the best academic outcomes for ELLs (Genesee, Lindholm-Leary, Saunders, & Christian, 2005; Gersten & Baker, 2000; Slavin & Cheung, 2005). However, Slavin and Cheung’s (2005) synthesis included ELD as part of the “English immersion” category, thus masking the distinction of interest in the current study. Gersten and Baker (2000) focused on components that they determined should be included in ELD instruction, rather than on the overall effects of ELD versus mainstream English classes. Thus, these two reviews do not directly address the question of the effects of instruction with and without ELD support. As part of their review, Genesee et al. (2005) compared several outcomes related to instruction that either incorporates ELD or does not but cautioned that the “actual research base is scant and is composed mostly of correlational studies” (p. 377).

Following the publication of these review articles, several studies used observational data and tried to condition or match on observable variables to estimate the effect of ELD services. First, Callahan, Wilkinson, Muller, and Frisco (2009) used Add Health data and a propensity score matching technique to match students in ESL settings with observationally similar students in mainstream English classrooms during high school. The findings suggested that students in the ESL courses in schools with high concentrations of immigrant students—schools similar to those in the current study—enrolled in more advanced mathematics and science courses, had higher grade point averages, and failed fewer courses. In a second study, Callahan et al. (2010) again used a propensity score matching strategy and concluded that recent immigrants and low-English-proficiency ELLs may benefit somewhat from ESL, while higher English proficiency or long-term ELLs are likely harmed by ESL services in high school. Finally, Flores et al. (2009) used regression analysis to conclude that reclassification is on average beneficial, and thus students should be reclassified sooner. However, the methods used in these studies do not allow us to disentangle the true effects of reclassification from unobservable factors that influence the reclassification decision (e.g., child motivation, a teacher’s beliefs about a child’s likelihood of success in ELD vs. non-ELD environments). In contrast to these studies, Matsudaira (2009) used an approach similar to the one discussed and implemented in this study and estimated that the effect of providing the 1st year of bilingual education or ESL to students who would otherwise be termed initially fluent English proficient is minimal; however, this study and the current study differ substantially in their conceptual designs.

The way in which schools organize to deliver ELD instruction is often cited as promoting failure of ELLs and limiting their academic opportunities. In providing ELD classes to ELLs, schools (particularly secondary schools) tend to organize
in between-classroom ability groups (or “tracks”), creating a situation in which ELLs are physically separated from non-ELLs (e.g., Katz, 1999). Often, cross-track movement is difficult (see Oakes, 2005, for a general discussion of cross-track movement). For secondary school ELLs, ELD courses supplement or replace mainstream English language arts—and often mathematics, science, and social studies—courses, constituting tracks unto themselves (Valdés, 1998). Such secondary school tracks consisting of ELD or ESL courses have been termed a “dead-end path” for ELLs (Gutiérrez, 2005; Valdés, 1998). Gutiérrez (2005) recounted a senior district administrator saying, “The ESL track is a kind of purgatory, a holding place to nowhere.” Perhaps this criticism is justified, as many ELLs begin their elementary years in ELD and never exit by the time they complete high school (Ruiz-de-Velasco & Fix, 2000); and as the highest performing ELLs become reclassified as R-FEPs, the lowest performing students become concentrated in the ELD tracks (Abedi, 2004; Gándara et al., 2003; Gutiérrez, 2005; Valdés, 2001). Yet there is no guarantee that reclassification alone will shift ELLs into a much better track: Parrish et al. (2006) quoted an administrator who questioned what the practical difference is if a student is “an [English learner] in an ELD-style program or if they’re sitting at the lowest-level of the English-only classes, rock-bottom and failing.” The administrator continued, “[In this district] they’re more likely to get good instruction by still being ELL and having a CLAD-certified teacher” (§ 5, p. 23).

Highlighting the lack of options for secondary school ELLs, Gándara et al. (2003) described course schedules of primarily ELD versions of core subjects and physically education courses as place fillers. The overwhelming majority of these courses did not fulfill college-preparatory requirements, whereas the majority of courses taken by English-only students did. Moreover, Callahan (2005) found that 98% of the ELLs in her sample did not enroll in enough college-preparatory courses to apply for entry into California’s 4-year state colleges and universities and thus concluded that track placement was limiting opportunities for ELLs beyond high school. Note, however, that much of the criticism for the high school ELD track has been based on lower level ELLs’ performing far below the thresholds for reclassification, whereas in the current study, I compare higher level ELLs with barely R-FEPs at the policy-specified threshold to determine whether reclassification itself confers new opportunities.

**Focus on Reclassification as a Cause, With the Aim of Testing if the Transition Is Smooth**

Since instructional services and settings change because of reclassification status, reclassification should be viewed as a cause potentially affecting academic outcomes, rather than as an innocuous yardstick by which to measure the effectiveness of programs. Consider that although reclassification could have effects, it ideally should not affect outcomes at the assessment-based transition point if that transition point from ELL to R-FEP status is appropriate. No longer are we looking for “positive” effects, because such findings imply that students have been held back for too long. Rather, null effects indicate that a district has set the test score-based transition point at an appropriate level given the available instructional options. Importantly, anything other than a null effect implies that a better instructional program exists and is currently in use for otherwise identical students; that is, if there is a positive effect of reclassification, then students just below the current threshold could benefit from reclassification, while a negative effect implies that students just above the threshold could benefit from not being reclassified. A null effect implies that the transition between the currently used instructional programs is appropriate, but it does not rule out the possibility that a better unused instructional program exists (e.g., a curriculum not currently used in the district).

To illustrate this logic, the three panels of Figure 1 represent potential relationships between English proficiency (on the x-axis) and next-year CST-ELA scores (on the y-axis) given whether the student is instructed in the reclassified (solid line) or nonreclassified (dashed line) setting in three different scenarios (of an infinite number of potential scenarios; panels I, II, and III). In each panel, points a, b, and c represent different potential thresholds along the English-proficiency score that policy makers could conceivably set for reclassification, such that students above any chosen threshold become R-FEPs (and switch to the
reclassified setting) and those below it remain ELLs (and remain in the nonreclassified setting). The goal of the district, then, is to establish a curriculum and corresponding threshold for reclassification that ensures a smooth transition from ELL to R-FEP status.

Panel I shows the only scenario of the three in which an ideal transition point exists; fortunately, though, this is a very plausible scenario in which the nonreclassified setting (including ELD services) is more beneficial to students of lower English proficiency, and the benefits diminish.
and ultimately become negative at higher levels of English proficiency, a pattern compatible with recent evidence from Callahan et al. (2010). Thus, in panel I, policy makers should aim to set the threshold at point \(a\), at which the achievement profiles for students in the reclassified and non-reclassified settings cross (i.e., where the marginal effect of ELD services is zero). Setting the threshold at point \(a\) amounts to reclassifying ELLs when they are still better served in the nonreclassified environment (e.g., before they are ready to move into an English-only classroom without ELD support) and scoring higher than they would if they were reclassified. Alternatively, setting the threshold at point \(c\), at which reclassified students perform much higher than otherwise identical students in nonreclassified settings, means that the students who remain in nonreclassified settings between points \(b\) and \(c\) are learning less in those settings than they would in the reclassified setting. In other words, students at point \(c\) should be reclassified because they benefit more in the reclassified setting than in the nonreclassified setting, but the threshold for reclassification should be set at point \(b\) to reclassify any student at or above point \(b\) (i.e., the group that would benefit from reclassification).

Panel II illustrates a scenario in which the marginal gains to switching to the reclassified setting are constant throughout the English proficiency continuum. That is, if policy makers set the threshold for reclassification at point \(c\), they would observe positive effects of reclassification. In response, they would likely want to reclassify more students by lowering the threshold to point \(b\). Again, they would observe positive effects of the same magnitude and again lower the threshold to point \(a\). Here, this resetting of the threshold would iterate until all students were reclassified regardless of English proficiency. Such a scenario might be found if the nonreclassified setting was characterized by a watered-down ELD version of the mainstream English language arts curriculum and yet the students were capable of learning more in the mainstream English language arts classes.

Finally, panel III presents a scenario in which switching to the reclassified setting is marked by increasing (rather than decreasing) marginal losses at higher levels of English proficiency. Assume that policy makers set the threshold at point \(a\) and observe a negative effect of reclassification. The likely response is to raise the threshold to point \(b\), to reclassify fewer students, so that fewer students experience negative effects. Although fewer students would indeed experience negative effects (a desirable result of moving the threshold from point \(a\) to point \(b\)), the marginal loss to reclassification grew larger, a counterintuitive result if one believes the defining feature of the ELL classroom (relative to the mainstream classroom) is ELD and that ELD accomplishes its desired objective. Such a situation must prompt one to question what truly characterizes the nonreclassified educational setting and/or whether ELD instruction is achieving its principal objective. If this scenario occurred, one might also want to question which mechanisms contribute to this pattern. One possibility could be an accumulated disadvantage due to a history of watered-down curricula, which leaves ELLs unprepared to fully engage with the academic material once reclassified (cf. Gutiérrez, 2005; Valdés, 2001). Although the short-term solution may be to raise the policy threshold higher, a longer term solution would likely involve a deeper examination of the processes that contributed to this pattern.

Importantly, note that the statistical design of this study (discussed next) makes no assumption about which scenario holds. Indeed, multiple scenarios can exist simultaneously (i.e., different scenarios for different types of students), a topic to which I return in the discussion section.

### Methods

**Sources of Selection Bias in the Reclassification Process and How to Remove That Bias in an Evaluation**

For many ELLs in California’s public schools, the reclassification process begins when they attain a combination of required scores on five assessments: the CST-ELA, the overall CELDT, and the three CELDT subcomponents (reading, writing, and listening/speaking). In addition to attaining a score of 300 on the CST-ELA, this district requires students to pass the overall CELDT and each of its subcomponents at the level of “early advanced.” The district office informs the school of attendance when an ELL has successfully attained the cut scores for these five assessments; following this, the teacher informally rates...
the student’s English proficiency and academic ability (using a variant on the Student Oral Language Observation Matrix, as well as course grades and impressions). Then, a small group of school faculty members, usually at least the child’s teacher and the ELL coordinator, meet to discuss whether the child should be reclassified as R-FEP or remain as ELL.

Because failing to meet even one of the five tests can be enough to stand in the way of a student’s reclassification, we can think of the minimum value of the set of five tests as the “gatekeeper” or “binding” test for that student. In fact, a new variable can be created to reflect the binding score, thereby reducing the dimensionality of the problem from five test scores to one composite score (Martorell, 2005; Reardon & Robinson, in press): Each child $i$’s minimum (or binding) score $M$ is a function of his or her prior-year ($t – 1$) assessments: $M_{i,t-1} = \min(CST-ELA_{i,t-1}, \text{overall CELDT}_{i,t-1}, \text{reading CELDT}_{i,t-1}, \text{writing CELDT}_{i,t-1}, \text{list/sp CELDT}_{i,t-1})$.

In elementary schools in the district analyzed, for example, students falling short of the required score on any one of these five measures have virtually no chance of reclassification; however, students just meeting the final (of the five) assessment requirements have about 0.79 greater probability of reclassification than students scoring just below the threshold on the final of these tests (see Appendix B, Table B1). Note that the change in probability of reclassification for attaining the final assessment threshold is 0.79 (not 1) because other, endogenous factors (e.g., teacher evaluations, parent opinions) also affect reclassification decisions. Therefore, the method used for obtaining unbiased effects of reclassification must use the much improved chance of reclassification for students just attaining the required cut scores but remove the nonrandom factors affecting both reclassification decisions and educational outcomes.

A simple comparison of average current achievement differences between R-FEPs and ELLs conflates reclassification effects with ability, because reclassification is based in part on prior achievement, and prior achievement is a strong predictor of current achievement. Less obvious, statistically conditioning on prior achievement will still result in biased estimates. This is because reclassification is based partly on achievement criteria and partly on teacher and parent decisions. For instance, if two students both score above the reclassification criteria, but only one is reclassified, we must ask, “Why didn’t the other student get reclassified too?” The answer is likely that the teacher thought the other student would not succeed in the mainstream English classroom without ELD support services. An ordinary least squares (OLS) regression of achievement on reclassification status, even conditioning on prior achievement, would not account for this factor that is unknown to the researcher.

RD-IV (also known as “fuzzy RD”; see Angrist & Lavy, 1999; Trochim, 1984) can estimate the effect of reclassification only when there is a sudden, and known, change in the likelihood of reclassification. For example, the left panel of Figure 2 presents the proportion of students in Grades 4 to 6 who are reclassified at different points along the binding score dimension. Students scoring just below the binding cut score have virtually no chance of being reclassified, but students scoring just at the threshold are reclassified about 80% of the time. There is no reason to think that about 80% of students just barely meeting the final assessment criteria are motivated or ready for reclassification, yet almost 0% of the students just barely failing to meet the criteria would also be deemed ready for reclassification (and possess the same average motivation level); however, they are not reclassified, because they did not meet the full set of assessment criteria. Therefore, the method used in this study capitalizes on this sudden exogenous shift in the likelihood of being reclassified at the threshold, thereby removing any selection bias.

To be valid, the “instrument” (here, achieving the final cut score) must (a) predict reclassification status and (b) not predict the outcome of interest (e.g., current-year CST-ELA scores) other than through its effect on the reclassification status, conditional on the observed factors included in the model (e.g., prior achievement). With respect to the first condition, Figure 2 and Table B1 provide convincing evidence that achieving the five cut scores in year $t – 1$ predicts whether an ELL is reclassified in year $t$. With respect to the second condition,
passing the cut score should not predict achievement the following year, as long as a flexible function of prior-year achievement for all five assessments is included in the achievement equation, as in all equations below, and provided there is no manipulation of student test scores. Because the CST-ELA and most types of CELDT assessments are scored off site, manipulation is very unlikely.8 (A more detailed discussion of checks for manipulation can be found in Appendix B.)

Binding-Score RD Design With IVs

The method for estimating the causal effect of reclassification on the type of student who would be reclassified if he or she attained the policy-based assessment thresholds can be decomposed into two equations. The first-stage equation isolates the contribution of the known mechanism for determining whether students are reclassified (i.e., the portion associated with attaining the final threshold) from the portion of reclassification on the basis of unobserved factors. Specifically, Stage 1 predicts reclassification status \( R_{i,t} \), by linear and quadratic terms for the binding score; a set of interactions between attaining the cut score on all five assessments and the linear and quadratic continuous binding score; student demographics (i.e., gender, race, special education status, free and reduced-price lunch status); indicators for which assessment (e.g., prior-year CST-ELA) contributes the binding score for each student; and an indicator variable for attaining the cut scores, where \( C = 1 \) if all five cut scores have been attained (i.e., if the binding score threshold is met), and \( C = 0 \) if one or more cut scores have not been attained (i.e., if the binding score threshold is not met). The first-stage equation is

\[
R_{i,t} = f(M_{i,t-1}, C_{i,t-1}) + \alpha(C_{i,t-1}) + \mathbf{X}'_{i,t}\beta + u_{i,t}.
\]

Because the minimum score is the binding constraint, then (provided the other scores vary smoothly across the minimum threshold) only a flexible function \( f(\cdot) \) of the minimum score \( M \) is required to obtain an unbiased estimate (see Martorell, 2005; Reardon & Robinson, in press). To improve the precision of the estimates, linear terms for each of the five assessments and student demographics are also included (as mentioned in the above paragraph and represented in equations by the vector \( \mathbf{X} \)). Note here that alternative specifications (not presented) omitted the variables in \( \mathbf{X} \), and similar results (with larger standard errors) were obtained, as expected.

In the second-stage equation, I use the predicted value of \( R \) (i.e., \( \hat{R} \)), predicted from equation 1, instead of \( R \) to predict an outcome \( Y \) (e.g., CST-ELA, attendance, college-preparatory credits) for child \( i \) in year \( t \) (or \( t + 1 \)). That is,

\[
Y_{i,t} = f(M_{i,t-1}) + \delta(\hat{R}_{i,t}) + \mathbf{X}'_{i,t}\theta + \epsilon_{i,t}.
\]

After conditioning on the variables (other than the instrument \( C \)) in the Stage 1 equation, \( \hat{R} \) contains only the portion of \( R \) that is determined by attaining the cut score; therefore, \( \hat{R} \) is assumed to be uncorrelated with \( \epsilon \). Thus, \( \delta \) in equation 2 is the unbiased local effect of reclassification on the outcome of interest. This local effect is only generalizable to students who \( (a) \) would be
reclassified only if they achieved all five cut scores (the IV restriction to generalizations) and (b) are close to passing the final cut score (the RD restriction to generalizations). That is, the estimate of \( \delta \) is not necessarily generalizable to students (a) who would be (or would not be) reclassified regardless of whether they attained the thresholds or (b) who are at other points in the prior-year test-score space (say, for a student scoring below 290 on the prior-year CST-ELA). However, because the goal is to assess the appropriateness of the level at which the district chose to set its reclassification threshold to guide teachers’ decisions, the estimated effect of reclassification from equation 2 has important and direct policy implications.

Other factors associated with achievement (e.g., motivation) are assumed to vary smoothly across the cut score. For example, we assume students who attain all CELDT thresholds and score 299 points on the CST-ELA are just as motivated as students who students attain all CELDT thresholds and score 300 points on the CST-ELA. This is a fairly innocuous assumption, because teachers and students are very unlikely to manipulate (or even be able to manipulate) on which side of the cut score a student falls. Assuming this, the same proportion of highly motivated students should be on either side of the cut score. However, only the group of students above the cut score will have a high likelihood of being reclassified. Therefore, the estimate obtained from this previously described process answers the question, Given the students who would be reclassified if they attained the threshold, what is the effect of reclassification for these students near the threshold? This tells us the effect of reclassification for this subpopulation of students, but it also tells us whether the district-set threshold is placed at an appropriate level given the currently available instructional options.

Data

The rich data sets for the analyses come from a large urban school district in California, with ELLs constituting about one third of the district’s student population. Variables include school year, grade level, race, gender, special education status, English learner status, attendance, grade point average, CELDT scores, and CST scores. The data are longitudinal, student-level observations following children as they progress through school, and thus the previously mentioned variables are available at multiple time points. Restricting the data set to students with three consecutive years of valid test score data and valid information on English learner status, there are 166,245 student-year observations. The availability of multiple years (from 2001–2002 to 2006–2007) and cohorts of data, spanning 3rd through 11th grades, permits tracking individual students while accounting for prior achievement and changing statuses (e.g., changing from ELL to R-FEP status). Additionally, course history files allow examining changes in advanced coursework (i.e., college-preparatory courses) that occur when a student is reclassified.

Asian (primarily Hmong) and Hispanic ELLs were retained for analyses because they had ample numbers of students; White, Black, and other ELLs, together accounting for less than 3% of the district’s ELL population, were not retained for analyses. Native English speakers and language minority students who were initially deemed fluent English proficient are not retained for analyses because, by definition, they never experience reclassification. This left 39,736 potential observations.

Because the goal of this study is to estimate the effect of reclassification for students on the cusp of being reclassified, and therefore to assess the district’s chosen achievement level for transitioning ELLs to R-FEPs, students far from attaining the criteria are dropped from the analyses. Specifically, only students within 1 standard deviation of the threshold on the final (binding) score were retained for analysis. This reduced the total observations from 39,736 to 22,827. Alternative analytic samples are discussed (and their results are presented) in Appendix B, but the conclusions remain the same. Thus, the main text proceeds with the 1 standard deviation analytic sample.

Table 1 presents descriptive statistics for the analytic sample. Male and female students are represented in almost equal proportions in all grade levels, though there is a tendency for slightly larger proportions of female students in the earlier grades. Asian students represent larger proportions of the sample as grade level rises, from 36% in 4th grade to 51% in 10th grade. The proportion of students receiving free or reduced-price lunch is uniformly high, at 99% in all elementary and middle school grades, but drops to around 90% in high school. The proportion of ELLs identified for special
Reclassification rates in the sample reach a high of 26% in 7th and 8th grades, increasing from a low of 16% in 4th grade. Average attendance rates for the sample are high, around 97% across grade levels. Grade point average (in the fall before the reclassification decision) is available for middle and high school students, and is 2.67 (out of 4.0) in middle school and 2.28 to 2.22 in high school.

CST-ELA and all CELDT scores are the average scores prior to the reclassification decision, standardized ($SD = 1$) and recentered around their respective cut scores (i.e., a negative mean indicates that the sample scored below that cut score on average). As grade level increases, there is an interesting shift in the CST-ELA and CELDT means: Mean CST-ELA decreases, indicating lower English reading and language arts skills (relative to the grade-level standards), but mean CELDT increases, suggesting better basic communication skills in English as grade level increases.12

Which Tests Contribute the “Binding Score”? As previously stated, the binding score—the minimum of five tests, standardized and recentered around their respective cut scores—is calculated separately for each child. That is, the CST-ELA threshold may be the binding score for one child, but the CELDT reading score may provide the binding score for another child. Figure 3 presents the proportion of students in the sample who have certain tests as their binding scores, by grade level. This information is relevant for several reasons. First, it identifies which tests serve as the “gatekeepers” to reclassification in different grade levels and documents how these gatekeepers change with grade level. Second, when combined with Table 1 and Appendix A, it provides rich context. For example, looking at the sample of high school students, Table 1 states the average achievement on the CST-ELA is just below the cut score in 9th grade (at –0.02) and less than 0.10

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Grade level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Male</td>
<td>0.44</td>
</tr>
<tr>
<td>Asian</td>
<td>0.36</td>
</tr>
<tr>
<td>FRPL</td>
<td>0.99</td>
</tr>
<tr>
<td>Special education</td>
<td>0.02</td>
</tr>
<tr>
<td>Reclassified</td>
<td>0.16</td>
</tr>
<tr>
<td>Attendance</td>
<td>97.59</td>
</tr>
<tr>
<td>(2.89)</td>
<td>(3.18)</td>
</tr>
<tr>
<td>Fall GPA</td>
<td>NA</td>
</tr>
<tr>
<td>(0.89)</td>
<td>(0.86)</td>
</tr>
<tr>
<td>CST-ELA</td>
<td>0.25</td>
</tr>
<tr>
<td>(0.79)</td>
<td>(0.85)</td>
</tr>
<tr>
<td>CELDT</td>
<td>0.15</td>
</tr>
<tr>
<td>(0.65)</td>
<td>(0.67)</td>
</tr>
<tr>
<td>CELDT reading</td>
<td>–0.12</td>
</tr>
<tr>
<td>(0.57)</td>
<td>(0.66)</td>
</tr>
<tr>
<td>CELDT writing</td>
<td>0.09</td>
</tr>
<tr>
<td>(0.60)</td>
<td>(0.62)</td>
</tr>
<tr>
<td>CELDT listening/speaking</td>
<td>0.27</td>
</tr>
<tr>
<td>(0.82)</td>
<td>(0.81)</td>
</tr>
<tr>
<td>n</td>
<td>3,292</td>
</tr>
</tbody>
</table>

Note. CELDT = California English Language Development Test; CST-ELA = California Standards Test of English Language Arts; FRPL = free and reduced-price lunch; GPA = grade point average; NA = not available. The analytic sample includes students whose binding scores are within 1 standard deviation of the cut score. Standard deviations appear in parentheses below means of continuous variables. GPA is on a scale ranging from 0.0 to 4.0. CST-ELA and all CELDT scores were standardized in the full sample and then recentered around their respective cut scores.
standard deviations above the cut score in 10th grade (at 0.09). Compare this with the CELDT reading test, for which students are 0.57 and 0.77 standard deviations above the cut score in 9th and 10th grades, respectively. This suggests that the CST-ELA is a relatively harder test for high school students in the sample. Figure 3 corroborates this by illustrating that high school ELLs are more likely to remain ELLs because of their CST-ELA scores ($\geq 46\%$) rather than any one CELDT score ($\leq 26\%$). Specifically, Figure 3 gives a categorical picture of which test is binding, and Table 1 gives a sense of the magnitude of how far above (or below) the cut score students are on average.

Because reclassification occurs a few months prior to the next administration of CSTs, “year of” outcomes models will capture the proximal effects of reclassification a few months after reclassification, while the outcomes models referred to as “year after” will reflect the effects roughly 15 months after reclassification occurred. All estimates are in standard deviations on the sample of ELL and R-FEP students in the school district and can thus be interpreted as effect sizes.

Raw differences. The first column gives raw differences, showing that R-FEPs outscore ELLs by large and significant margins in all school levels analyzed. For example, in elementary school (Grades 4 to 6), R-FEPs outperform ELLs by more than 0.9 standard deviations on the current-year CST-ELA. Similar differences are found for middle and high school, though the raw gap is somewhat smaller at about 0.75 to 0.8 standard deviations in later grades; this reduction primarily reflects a shift in sample density along the prior-year score (closer to the cut score). These estimates should not be interpreted causally.

## Results

### Effects on CST-ELA Achievement

Table 2 presents estimated differences between R-FEPs and ELLs on the CST-ELA in different school levels (i.e., elementary, middle, or high school), both for the year in which reclassification occurs and 1 year after students are reclassified.
The second set of estimates uses OLS regression, with each level-specific model conditioning on gender, special education status, free and reduced-price lunch status, five tests of prior achievement, linear and quadratic terms for the binding score, indicators for which test contributes the binding score, and year fixed effects. The rather extensive set of covariates accounts for the bulk of the achievement difference between R-FEPs and ELLs, yet the differences remain positive and significant in all grade levels. One year after reclassification, the positive estimates persist, with no indication of fading effects. Although these models condition on prior achievement and other demographic characteristics, they cannot account for unobserved factors, so there may be bias in these estimates.

**OLS regression estimated at the cut score.** Before moving to the binding-score RD-IV estimates, I present one more set of OLS estimates. The prior set of OLS results estimated the reclassification difference over the entire sample, which is the conventional approach (see, e.g., Flores et al., 2009). Because no students should be reclassified without attaining all five assessment thresholds, these OLS models base their estimation on extrapolating relationships beyond points that the data can support. Therefore, in addition to the selection bias issue, these OLS estimates hinge on the tenuous assumption that the relationship between prior achievement and current achievement is the same for ELLs and R-FEPs (even though there cannot be any R-FEPs at the lower end of the achievement distribution). A better (though still prone to selection bias) set of OLS regressions estimates the relationship between reclassification and achievement at the binding-score cut score; this third set of estimates in Table 2 is the OLS analogue to the binding-score RD-IV.

### Table 2

<table>
<thead>
<tr>
<th>School level</th>
<th>Year of</th>
<th>Year after</th>
<th>Year of</th>
<th>Year after</th>
<th>Year of</th>
<th>Year after</th>
<th>Year of</th>
<th>Year after</th>
<th>Year of</th>
<th>Year after</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>0.94***</td>
<td>0.84***</td>
<td>0.12***</td>
<td>0.09***</td>
<td>0.10***</td>
<td>0.08**</td>
<td>0.06</td>
<td>0.03</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Middle</td>
<td>0.78***</td>
<td>0.77***</td>
<td>0.11***</td>
<td>0.12***</td>
<td>0.09**</td>
<td>0.12***</td>
<td>0.10</td>
<td>0.04</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>Elementary/middle</td>
<td>0.86***</td>
<td>0.80***</td>
<td>0.11***</td>
<td>0.11***</td>
<td>0.10***</td>
<td>0.10***</td>
<td>0.07†</td>
<td>0.03</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>High</td>
<td>0.75***</td>
<td>0.70***</td>
<td>0.17***</td>
<td>0.21***</td>
<td>0.13**</td>
<td>0.16***</td>
<td>−0.13</td>
<td>−0.31*</td>
<td>−0.11</td>
<td>−0.37*</td>
</tr>
</tbody>
</table>

### Note

ELL = English language learner; OLS = ordinary least squares; RD-IV = regression discontinuity with instrumental variables; R-FEP = reclassified fluent English proficient. Estimates are presented in pooled standard deviation units and can be interpreted as effect sizes. Robust standard errors appear in parentheses. Sample sizes for elementary (Grades 4–6), middle (Grades 7 and 8), and high (Grades 9 and 10) school were 11,134, 6,907, and 4,786, respectively. The “elementary/middle” models pool the elementary and middle samples (n = 18,041), providing a single estimate per model; these models also include an indicator for school organization level. All models (except the unadjusted model) include covariates for gender, special education status, race (Asian or Hispanic), lunch program status, year fixed effects, linear terms for each of the five achievement tests in reclassification decision, linear and quadratic terms for the binding-score assessment, and a series of indicator variables for which test contributed the binding score. The OLS model evaluated at the cut score adds interactions between reclassification and the linear and quadratic binding-score assessment variables. The parametric RD-IV model adds interactions between being above the cut score and the linear and quadratic binding-score assessment variables and uses attaining the binding-score threshold as the instrument for reclassification. The nonparametric RD-IV uses Imbens’s (2009) stata rdob ado file.

†10%. *5%. **1%. ***0.1%.
estimates are statistically indistinguishable from the prior set of OLS estimates, though the new estimates are slightly less positive than the original OLS estimates. Again, though, these estimates are susceptible to selection bias and should not be interpreted causally.

**Binding-score RD-IV.** In the year of reclassification, all point estimates are statistically indistinguishable from zero, a contrast with the significant positive OLS estimates. Moreover, the elementary and middle school estimates in the year following reclassification continue to approach zero and remain nonsignificant, suggesting that reclassification, given these particular thresholds and instructional practices, does not have any immediate or year-after effect on students in Grades 4 through 8. Thus, there is evidence of a smooth transition in these grades, suggesting that the thresholds are appropriate. However, the standard errors of the RD-IV estimates are far larger than those of the corresponding OLS estimates, because of the use of RD and IV, two methods known to have reduced power (for RD, see Schochet, 2009; for IV, see Jo, 2002). The OLS estimates for the elementary and middle school samples fall within the larger confidence interval of the binding-score RD-IV estimates, and thus the OLS estimates are not inconsistent with the RD-IV results. To improve precision in the RD-IV estimation, the elementary and middle school samples (which yielded similar estimates) were pooled for a secondary analysis. The result can be seen in the third row of Table 2, where there is weak suggestive evidence of a positive year-of effect through eighth grade in the parametric model only. Generally speaking, though, the elementary and middle school RD-IV suggests that there is no strong evidence of a reclassification effect, particularly given the attenuating year-after effect. Thus, the transition point seems suitable.

Unlike the earlier grades, the year-after effect in high school does not attenuate but rather reaches a significant year-after effect of –0.31 standard deviations in the parametric model and –0.37 standard deviations in the nonparametric model. (Figure 4 shows the reduced-form binding-score RD for the year-after intent-to-treat effect in Grades 9 and 10.) Also different from the earlier grades, the corresponding OLS estimate (+0.16 standard deviations, \( p < .001 \)) is significant in the opposite direction and is consequently not contained within the RD-IV confidence interval for high school. Thus, the RD-IV results imply that the nonreclassified setting was more beneficial than the reclassified setting for fostering English language arts achievement in the students at the threshold; this is the direct opposite of the conclusion one would reach on the basis of the conventional approach.

![Reduced-form binding-score regression discontinuity (RD) for Grades 9 and 10. CST-ELA = California Standards Test of English Language Arts.](http://eepa.aera.net)
To assess the robustness of the binding-score RD-IV findings, several checks were performed. First, parametric and nonparametric models were run, which resulted in similar conclusions. In addition, Appendix B further illustrates that the results are robust to sample selection procedures and are not the product of an ill-fitting parameterization (see Table 2 and Appendix B, Figure B1). Appendix B also contains details on the smoothness of \( X \) variables at the binding-score cut score, RD-IV analyses by test contributing the binding score, and checks for discontinuities in reduced-form equations at placebo cut scores (see Appendix B, Table B3). All of the evidence suggests that the observed negative year-after effects in high school are indeed due to reclassification, not to condition manipulation or selection bias, anomalous data, or spurious relationships.

Why do we see a bigger discrepancy between the OLS and RD-IV estimates in high school than in earlier grades? One possible reason is the degree of compliance with reclassifying students if and only if they attain the required cut scores: The greater the compliance to the cut score criterion, the less room for selection to bias the OLS estimates. Through Grade 8, virtually none of the students falling short of the final cut score were reclassified, and about 80% just attaining the final cut score were, meaning that compliance was high and attainment of the threshold was by far the dominant factor in reclassification decisions. Thus, there was little room for an omitted variable (e.g., teacher’s perception of the student’s ability) to bias the OLS estimates. Compared with these lower grades, in high school a larger percentage of students just barely failing to attain the final cut score were reclassified (about 12%), and a lower percentage of students barely attaining the final score were reclassified (about 60%), both contributing to a drop in compliance. Thus, attaining the final cut score plays less of a role in reclassification decisions, increasing the possibility for unobserved factors to bias OLS estimates. In Grades 9 and 10, in particular, the concern is that the conventional OLS approach would generate estimated “effects” of reclassification that were higher than the true effects (approximated by RD-IV) because compliance to the cut score criterion is low, which may be due to teachers’ selecting good students to reclassify. Indeed, in high school, the OLS estimates are more positive than the binding-score RD-IV estimates in the years of and following reclassification, resulting in very different conclusions (i.e., opposite signs of the estimated effects).

This discrepancy between the OLS and binding-score RD-IV estimates in high school prompts an important question: Do the positive OLS estimates suggest that teachers and administrators are appropriately selecting the students who benefit from being reclassified? Not necessarily, if teachers have a tendency to reclassify students who would perform well regardless of instructional circumstances. This is because the OLS coefficient estimate on reclassification is a combination of (a) the RD-IV effect estimate (i.e., the effect on students who would be reclassified if and only if they attained the final threshold; the “compliers” in the terminology of Angrist & Imbens, 1995) and (b) the average achievement difference between students who would be reclassified regardless of whether they attained the threshold minus that of the students who would not be reclassified regardless of whether they attained the threshold. This second component biases the OLS estimate. Because there is no way to know if the group of “reclassified regardless” students scored higher than the “not reclassified regardless” students because of reclassification, in spite of reclassification, or regardless of reclassification (e.g., they were motivated to do well in any setting), we do not know the effect of teachers’ decisions regarding these students. What we do know, however (from the RD-IV estimates), is that for students whose teachers would reclassify them if they attained the threshold (and would not reclassify them if they did not attain the threshold), the year-after effect of reclassification in high school is negative. This suggests that the threshold that guides teachers’ decisions is not set at an appropriate point (i.e., is set too low) in high school.

**Effects on Course Taking and Attendance**

In addition to achievement effects, reclassification has been linked to improved access to advanced curricula (Callahan, 2005; Gandara et al., 2003; Gürtterrez, 2005). In the district analyzed here, one might think that reclassification would again result in improved access: In this district, ELLs can take a course only if it is taught by a CLAD- or BCLAD-certified teacher. If such teachers do not provide instruction in college-preparatory

---

281
courses (i.e., A–G courses in California), then ELLs will not be able to enroll in these courses. The OLS estimates are consistent with prior research suggesting reclassification results in improved opportunities (cf. Callahan, 2005). However, the binding-score RD-IV analyses revealed no evidence that reclassification affected either enrollment opportunities or successful completion of such courses, in any subject studied (English language arts, math, science, or overall college-preparatory credits). Thus, this pattern of results suggests that reclassification alone does not always open doors to A–G coursework; instead, students selected to be reclassified are also completing more A–G requirements, even after conditioning on prior achievement and English proficiency. Similarly, for the models using attendance as the outcome, OLS models found a positive effect, but binding-score RD-IV estimates revealed that reclassification has no effect on attendance. For parsimony, these results are not included in this article, but they can be provided by the author upon request.

The findings of this research should not be taken to mean that reclassification has no effect on curricular access or access to tertiary schooling in all cases. Instead, they should encourage us to scrutinize findings suggesting that there is an effect and ensure that these estimates are purged of selection bias; indeed, the OLS estimates in the current study (conditioning on five prior tests, other observable factors, and year fixed effects) found significant positive associations between R-FEP status and each outcome. But these OLS estimates do not account for unobserved factors related to reclassification, and the relationship disappeared when using the binding-score RD-IV.

Policy and Evaluation Implications

Reclassification of English learners is a hotly debated topic, as evidenced by the classic example of Proposition 227 proponents’ focus on low reclassification rates to claim that bilingual education was failing to promote English proficiency and thus rally support for the proposed law (see Grissom, 2004). Yet despite its ability to bolster a political campaign restricting bilingual education, little is known about the effects of reclassification (see Linquanti, 2001). Prior research on its effects has focused primarily on course-taking differences (Callahan, 2005; Callahan et al., 2009, 2010; Gándara et al., 2003; but see Flores et al., 2009); moreover, the results of these studies may be limited by their methodologies, which do not remove the influence of unobserved factors on reclassification decisions. The few studies that did explore achievement differences (e.g., Gándara et al., 2003; Grissom, 2004) often did not make any adjustments for prior achievement, and when they did make these adjustments (Flores et al., 2009), their estimates were still likely to be biased by reclassification on unobservable variables. Thus, prior research has not provided convincing estimates of the causal effect of reclassification on either course taking or achievement. In addition, the literature on reclassification tends to treat it as a yardstick by which to measure the effects of other factors or policies (e.g., Proposition 227). When it was recognized that reclassification itself can have effects, the assumption was that positive effects are desirable and negative effects are not.

The goals of this study were twofold. First, I argued that because reclassification is a change in instructional settings from one bundle of services to another, it can have effects on academic outcomes; however, if reclassification is timed properly, there should be no effect of this service change. Thus, what is desirable is a smooth transition from the nonreclassified setting to the reclassified setting, which would result in a null effect of reclassification (but see the discussion of the implications of null effects below). Second, I outlined a general methodological framework (i.e., using a binding-score RD-IV) that state and district policy makers can use to evaluate their own criteria and in the process discussed why conventional methods (e.g., OLS regression) will not suffice and may even lead us to the opposite conclusion. In the case of the particular district I chose to illustrate this general methodology, the results suggested that the chosen threshold for reclassification does not always provide null effects; that is, the existing reclassification policy is currently harming some students (in high school) in the district analyzed. This finding of a negative effect on subsequent-year English language arts scores in high school strongly suggests that we should not necessarily treat reclassification as a (positive) yardstick by which to measure the effectiveness of other programs or policies.
Before further discussion of the implications of the current findings, it is worthwhile to discuss some additional features of the analytic framework introduced in this article and to explain how this framework can be modified and extended to cases of systematic heterogeneity in the effects of reclassification at a threshold.

**Extensions to Differential Effects and a Set of Reclassification Thresholds**

Recall that Figure 1 presented three potential scenarios relating a student’s English proficiency to his or her subsequent-year achievement on an English language arts assessment, under two potential settings (the reclassified setting and the nonreclassified setting). In the analysis section, I estimated a series of average effects at the threshold for students whose reclassification status hinged on attaining the binding-score threshold. But it is possible that even at the same English proficiency level, different types of students or similar students instructed under different circumstances will have different responses to switching to the reclassified setting. If these moderating variables can be identified, the methods discussed in this article can be used to establish different thresholds in different contexts.

For example, Callahan et al. (2009) suggested that students in schools with smaller immigrant populations may be better served in a mainstream classroom without ESL, while students in schools with higher concentrations of immigrants may be better served by ESL. Policy makers could conceivably set a lower threshold in schools with lower concentrations of immigrants (so that these students spend less time in ESL settings) if indeed the effects of reclassification were moderated by the concentration of immigrants in a school. Other research suggests that length of time in the United States may moderate the effects of reclassification (Callahan et al., 2010; Conger, 2009; Ruiz-de-Velasco & Fix, 2000), and thus districts could establish different reclassification thresholds for students of different length residencies if there was empirical evidence that the effects of reclassification were moderated by the concentration of immigrants in a school. Other research suggests that length of time in the United States may moderate the effects of reclassification (Callahan et al., 2010; Conger, 2009; Ruiz-de-Velasco & Fix, 2000), and thus districts could establish different reclassification thresholds for students of different length residencies if there was empirical evidence that the effects of reclassification at the current threshold differed substantially by length of time in the United States. The actual estimation may involve stratifying observations along a dimension of interest (e.g., recent immigrants vs. longer term ELLs) and estimating the effects by each subgroup separately. The remainder would be a straightforward application of the methods discussed in this article.

Thus, the framework laid out in this article can be extended to explore whether there is empirical justification for establishing a set of reclassification thresholds, which vary by student attributes and/or contextual features (e.g., length of time in the United States). In practice, however, even if the attributes or features are identified and agreed upon, it may prove difficult to estimate precise effects if the number of students in the target group is small; as we have already seen, RD-IV yields less precise average estimates than the traditional OLS regression (see Table 2), and a smaller sample will only serve to reduce precision further. Even if a reliable set of heterogeneous thresholds cannot be established, the main empirical findings of this study illustrate that it is instructive to estimate an average effect at the chosen threshold. If the average effect is anything other than a null effect (as in high school, in the present data set), there is evidence for a serious concern and further investigation is warranted.

**The Implications of Null Effects**

Throughout this article, I have argued that finding anything other than a null effect is problematic because it suggests a poor transition, but this raises the question, Do null effects therefore imply that current practices are the “best” practices? No, a null effect does not automatically signify that the current practices are in any way “best” or even effective. Generally speaking, a null effect implies a better alternative may exist; a non-null effect implies a better alternative does exist. A null effect could be obtained because of insufficient power or if there were no differences between the services received in the nonreclassified and reclassified settings. A null effect also cannot tell us whether a third unobserved setting is preferred to the current nonreclassified and reclassified settings. That is, the RD-IV compares the existing settings, but just like the other methods mentioned, it cannot tell us if some unseen alternative is more desirable. Hence, the only thing we can conclude is that anything other than a null effect is indeed indicative that a better alternative exists (namely, the other instructional setting).
For example, in the particular data set used here, the analyses at the high school level demonstrated that at the current threshold, the nonreclassified setting resulted in significantly higher next-year CST-ELA scores than did the reclassified setting. This is a signal that action should be taken to modify the current criteria and/or instruction, as there is an existing setting (i.e., the current non-reclassified setting) that is more effective than another setting (i.e., the current reclassified setting). Consequently, the district may want to (a) keep the instructional options as is but move the threshold for reclassification higher, (b) keep the threshold as is, but change the instructional settings to better align the experiences (e.g., curricula, teacher qualifications, expectations, average peer achievement) of ELLs and R-FEPs near the cut score, (c) examine curricular differences in prior grades that could contribute to the observed negative effects in high school, or (d) some combination of these options. Any action taken, however, should be based on more in-depth qualitative knowledge of the students, processes that could have contributed to a difficult transition, district capacity, and feasible instructional alternatives. When null effects were found (as in elementary and middle school), policy makers may still choose to explore alternative strategies for ELL instruction to improve outcomes; however, the results of the binding-score RD-IV provide no guidance on which alternative (unobserved) instructional bundle or setting would lead to improved outcomes.

Concluding Thoughts

This study makes several key contributions to our knowledge of ELL reclassification. First, it shows that policy makers can greatly influence reclassification decisions by establishing assessment-based thresholds, even though other factors also affect the reclassification decision. Second, it argues that the resultant reclassification can have effects on student outcomes and illustrates that we should not assume these effects are positive. Third, it demonstrates that non-null effects of reclassification indicate a better instructional alternative exists for students at the current policy threshold. Fourth, it discusses that when assessing the effects of reclassification, conventional regression methods may produce biased estimates, particularly in cases of low compliance to cut score criteria. Fifth, it outlines and implements an alternative method for evaluating the effects of the assessment-based reclassification criteria. Reclassification is a highly complex subject, often unique to the circumstances of the district, its resources and instructional programs, its reclassification criteria, and its student population. While specific situations, and therefore effects, may be unique to localities, I suggest a common method for assessing these effects. In this article, I have outlined a general methodological framework for rigorously evaluating the effects of reclassification at a given threshold. Although applied here to a specific district for illustration purposes, the method can be adapted easily to study any locality with measurable thresholds as part of its reclassification criteria.

Appendix A
Adherence to Reclassifying Students Who Attain the Thresholds

Exploring the proportion of students reclassified who attained a given cut score tells us important information about the relative difficulty of the assessment threshold, as well as giving insights into the degree of teacher discretion in reclassification decisions. Table A1 presents the test passage rates, by grade level, for each of the five assessments used to determine reclassification. As grade level increases, fewer students attain the CST-ELA threshold, while more attain the CELDT thresholds. Essentially, passing the CST-ELA becomes a more binding constraint on reclassification as grade increases. This is also supported by Table A2, showing an increase in the proportion of students who were reclassified for passing the CST-ELA as grade level rises. Table A3 shows that very few or no students are reclassified if they failed to attain even one of the five required cut scores. Also worth noting is the drop in adherence to reclassifying students when...
TABLE A1
Test Passage Rates in the Analytic Sample, by Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>CST-ELA</th>
<th>CELDT overall</th>
<th>CELDT listening/speaking</th>
<th>CELDT reading</th>
<th>CELDT writing</th>
<th>All five tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.58</td>
<td>0.32</td>
<td>0.54</td>
<td>0.57</td>
<td>0.53</td>
<td>0.17</td>
</tr>
<tr>
<td>5</td>
<td>0.62</td>
<td>0.43</td>
<td>0.58</td>
<td>0.66</td>
<td>0.64</td>
<td>0.24</td>
</tr>
<tr>
<td>6</td>
<td>0.57</td>
<td>0.49</td>
<td>0.57</td>
<td>0.65</td>
<td>0.63</td>
<td>0.22</td>
</tr>
<tr>
<td>7</td>
<td>0.53</td>
<td>0.67</td>
<td>0.67</td>
<td>0.76</td>
<td>0.76</td>
<td>0.31</td>
</tr>
<tr>
<td>8</td>
<td>0.45</td>
<td>0.79</td>
<td>0.71</td>
<td>0.80</td>
<td>0.84</td>
<td>0.31</td>
</tr>
<tr>
<td>9</td>
<td>0.44</td>
<td>0.83</td>
<td>0.62</td>
<td>0.66</td>
<td>0.76</td>
<td>0.23</td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.88</td>
<td>0.67</td>
<td>0.76</td>
<td>0.83</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Note. CELDT = California English Language Development Test; CST-ELA = California Standards Test of English Language Arts.

TABLE A2
Reclassification Rates in the Analytic Sample, by Grade and Test(s) Passed

<table>
<thead>
<tr>
<th>Grade</th>
<th>CST-ELA</th>
<th>CELDT overall</th>
<th>CELDT listening/speaking</th>
<th>CELDT reading</th>
<th>CELDT writing</th>
<th>All five tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.28</td>
<td>0.30</td>
<td>0.28</td>
<td>0.49</td>
<td>0.30</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>0.34</td>
<td>0.33</td>
<td>0.31</td>
<td>0.48</td>
<td>0.36</td>
<td>0.85</td>
</tr>
<tr>
<td>6</td>
<td>0.35</td>
<td>0.31</td>
<td>0.30</td>
<td>0.39</td>
<td>0.34</td>
<td>0.83</td>
</tr>
<tr>
<td>7</td>
<td>0.49</td>
<td>0.34</td>
<td>0.34</td>
<td>0.38</td>
<td>0.38</td>
<td>0.81</td>
</tr>
<tr>
<td>8</td>
<td>0.57</td>
<td>0.31</td>
<td>0.31</td>
<td>0.33</td>
<td>0.36</td>
<td>0.80</td>
</tr>
<tr>
<td>9</td>
<td>0.42</td>
<td>0.24</td>
<td>0.25</td>
<td>0.22</td>
<td>0.28</td>
<td>0.67</td>
</tr>
<tr>
<td>10</td>
<td>0.43</td>
<td>0.26</td>
<td>0.28</td>
<td>0.24</td>
<td>0.32</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Note. CELDT = California English Language Development Test; CST-ELA = California Standards Test of English Language Arts.

TABLE A3
Reclassification Rates in the Analytic Sample, by Grade and Test(s) Not Passed

<table>
<thead>
<tr>
<th>Grade</th>
<th>CST-ELA</th>
<th>CELDT overall</th>
<th>CELDT listening/speaking</th>
<th>CELDT reading</th>
<th>CELDT writing</th>
<th>At least one test</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>9</td>
<td>0.00</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Note. CELDT = California English Language Development Test; CST-ELA = California Standards Test of English Language Arts.

they attain all five thresholds can also be seen in Figure 2, where the change in proportion reclassified at the threshold is smaller in high school than in elementary school. This drop signifies greater teacher discretion in the recategorization decision. While increased teacher discretion is arguably an improvement in the recategorization process, because teachers are monitoring a child’s progress throughout the year (rather than a test capturing a snapshot of achievement), teachers are more likely to reclassify highly able, or more motivated, students. The concern is that basing policy decisions on teachers’ selections of students most likely to achieve will lead to overestimated effects of recategorization. This drop in the recategorization rate for attaining all five cut scores in high school alerts us to look for greater discrepancies between the OLS and binding-score RD-IV model estimates in high school, as there may be more room for unobserved
factors to influence teachers’ reclassification decisions. In such cases, we would expect the OLS estimates to be more positive, suggesting that teachers reclassify the students most likely to score high, although this does not necessarily reflect an ability of teachers to select students most likely to benefit from being in a reclassified setting.

Appendix B

Robustness Checks

Specification checks for the parametric binding-score RD-IV. With RD models, once the selection mechanism is known, the primary requirement to obtaining unbiased estimates is fitting the regression model through the data (see, e.g., Trochim, 1984); this is no less important when using binding-score RD-IV or other forms of multiple rating-score RD (Reardon & Robinson, in press; Robinson & Reardon, 2009). As Robinson and Reardon (2009) suggested, a proper model can be chosen through visual inspection of the average residuals, calculated after any given RD model. If the model is appropriate, the average residuals should be close to zero throughout the sample domain. Therefore, to select the models presented in this article, I plotted the average residuals from several different binding-score RD-IV specifications. To illustrate this, Figure B1 presents plots of average residuals from four different binding-score RD-IV models used to estimate the year-after reclassification effect on CST-ELA scores in high school: (a) a linear term in the binding-score variable with cubic interactions between the binding-score variable and attaining the cut score, (b) quadratic with quadratic interactions, (c) cubic with quadratic interactions, and (d) cubic with cubic interactions. Figure B1 is presented for illustrative purposes, and because the year-after high school effects were significant in Table 2; similar visual inspections were done for all other grade levels and outcomes.

Most obvious in Figure B1 is the relatively poor fit of the linear with cubic interactions model. Other than this model, the residual averages did not display systematic overestimation or underestimation patterns. Importantly, regardless of which specification is chosen, the estimated year-after effects of reclassification are virtually unchanged. The quadratic model with quadratic interactions was chosen as the final model (with estimates presented in Table 2) because it used the fewest parameters (leading to improved precision), did not show worrisome systematic patterns of nonzero average residuals, and did not demonstrate any meaningful difference between its estimated effects and those of less parsimonious models. Its estimated year-after effect in high school was $-0.31$ ($SE = 0.16$); the models with cubic terms and either quadratic or cubic interactions yielded year-after estimated effects of $-0.41$ ($SE = 0.24$), significant at the 10% level. In fact, the preferred model’s estimated effect is of a slightly smaller magnitude than the more specified models, though the difference is statistically indistinguishable.

Also of note, the root mean square error in all of these models is between 0.7 and 0.9. Therefore, the fact that in all the specifications shown in Figure B1, the difference in average residuals between students on one side of the cut score and students on the other side is about 0.02 (roughly, $1/45$ to $1/35$ the size of the root mean square error) suggests that all the specifications available are reasonably good.

Alternative analytic samples. In addition to these specification checks on functional form modeling, two alternative analytic samples were used. The first restricted the sample further, to students within 0.5 standard deviations of the cut score, yielding a total sample (across all grades) of 12,672. The second expanded the sample to students 2 standard deviations from the cut score, resulting in a sample of 35,196 students. Across all analytic samples, the point estimates were very similar. Larger standard errors resulted from the models with few observations, and therefore, the negative estimate on year-after CST-ELA scores in high school was not significant in the 0.5 standard deviation sample, but it was significant in the 1 and 2 standard deviation samples (see Table B2). The stability of the estimated effect suggests the negative effect does exist in high school and is not a result of the analytic sample or modeling assumptions.

Smoothness of $X$ variables. Each $X$ variable (e.g., male, Asian, indicators for which test contributes the binding score) was examined for evidence of
FIGURE B1.  *Average residuals, by model fit of the binding-score variable.*

Note. CST-ELA = California Standards Test of English Language Arts. The numbers in parentheses are the model-specific number of parameters for the continuous binding-score variable (e.g., linear [1] with cubic interactions [3] has 4 continuous binding-score variable parameters).

TABLE B1

*Estimated Increase in Conditional Probability of Being Reclassified for Just Attaining the Final Threshold, by School Level and Analytic Sample*

<table>
<thead>
<tr>
<th>School level</th>
<th>Analytic sample (in standard deviations above or below binding cut score)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5 standard deviation</td>
</tr>
<tr>
<td>Elementary</td>
<td>Estimate</td>
</tr>
<tr>
<td></td>
<td>SE</td>
</tr>
<tr>
<td></td>
<td>n</td>
</tr>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Middle</td>
<td>Estimate</td>
</tr>
<tr>
<td></td>
<td>SE</td>
</tr>
<tr>
<td></td>
<td>n</td>
</tr>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td>High</td>
<td>Estimate</td>
</tr>
<tr>
<td></td>
<td>SE</td>
</tr>
<tr>
<td></td>
<td>n</td>
</tr>
<tr>
<td></td>
<td>F</td>
</tr>
</tbody>
</table>

*p = the *F* statistic on the instrument (attaining the cut score) in the first-stage regression predicting reclassification status, conditional on other covariates. The rule of thumb is that *F* statistics should generally be greater than 10 for the instrument to be valid (Stock, Wright, & Yogo, 2002; see also Stock & Yogo, 2005) when there is only one instrumental variable predicting one endogenous regressor (as in this study).
nonsmoothness across the cut score. Each demographic variable varied smoothly across the cut score. However, there was evidence that the CELDT listening/speaking more often provided the binding score for students just below the threshold in high school. Several checks (described in note 8) suggested this was not a concern for the estimation of the binding-score effects.

**Frontier RDs: Estimates stratified by test contributing the binding score.** In addition to the checks for manipulation in CELDT listening/speaking scores discussed in note 8, I inspected if the estimated effect of recategorization differed by which test contributed the binding score. That is, if manipulation resulted in significantly different types of students being reclassified, we should see very different effects of recategorization at the cut score of that specific test (i.e., CELDT listening/speaking) compared with the other tests’ cut scores. For these robustness checks, the data were stratified by test contributing the binding score (e.g., only students for whom the CST-ELA was the binding score), and nonparametric RD-IV models were estimated (i.e., a series of nonparametric “frontier RDs,” in the terminology of Reardon & Robinson, in press). The year-after reclassification effect point estimates at the thresholds of the CST-ELA (–0.34 standard deviations), CELDT writing (–0.32 standard deviations), and CELDT listening/speaking (–0.37 standard deviations) were similar, further suggesting that any manipulation in the CELDT listening/speaking test has a negligible impact on the effect estimates.

Note that stratifying the data by test contributing the binding score resulted in small sample sizes for the frontier RD-IV models of the CST-ELA ($n = 2,264$), CELDT writing ($n = 1,196$), and CELDT listening/speaking ($n = 1,090$). The sample sizes for CELDT reading ($n = 224$) and overall CELDT ($n = 32$) were too small to obtain reliable estimates. The small sample sizes of the individual frontier RDs resulted in low power; however, the homogeneity of point estimates provides evidence for combining the scores together in the binding-score RD-IV to improve power, as discussed by Reardon and Robinson (in press).

**Checks for discontinuities at placebo cut scores.** Finally, robustness checks at various placebo cut scores corroborate that the year-after reclassification effects observed in high school are not a

---

**TABLE B2**

*Binding-Score RD-IV Reclassification Estimates on English Language Arts Achievement Scores, by School Level, Analytic Sample, and Time Since Reclassification*

<table>
<thead>
<tr>
<th>School level</th>
<th>0.5 standard deviation analytic sample</th>
<th>1 standard deviation analytic sample</th>
<th>2 standard deviation analytic sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year of</td>
<td>Year after</td>
<td>Year of</td>
</tr>
<tr>
<td>Elementary</td>
<td>0.02</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Middle</td>
<td>0.17†</td>
<td>0.11</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>High</td>
<td>0.03</td>
<td>–0.39</td>
<td>–0.13</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.25)</td>
<td>(0.14)</td>
</tr>
<tr>
<td></td>
<td>2,783</td>
<td>2,783</td>
<td>4,786</td>
</tr>
</tbody>
</table>

*Note. RD-IV = regression discontinuity with instrumental variables. Estimates are presented in pooled standard deviation units and can be interpreted as effect sizes. Robust standard errors appear in parentheses. Sample sizes appear below standard errors. All binding-score RD-IV models include covariates for gender, special education status, race (Asian or Hispanic), lunch program status, year fixed effects, linear terms for each of the five achievement tests in recategorization decision, linear and quadratic terms for the binding-score assessment, and a series of indicator variables for which test contributed the binding score. The 2 standard deviation analytic sample models include cubic binding-score variables and an interaction between the cubic term and being above the cut score. †10%. *5%.*
spurious result that could be replicated at any point in the data (see Imbens & Lemieux, 2008). Table B3 shows that the only reduced-form equation resulting in a significant effect is at the true cut score, not any of the placebo cut scores. Thus, all of the evidence suggests that the observed negative year-after effects in high school are indeed due to reclassification, not to condition manipulation or selection bias or anomalous data.

Notes

1. Note that there are reliability and validity concerns when evaluating an ELL’s candidacy for reclassification on the basis of an assessment normed to an English-speaking population (see, e.g., Abedi, 2002, 2004). Although this may be undesirable and possibly unfair to ELLs, CST scores are nevertheless used for reclassification determination, and thus the analyses will be able to estimate causal effects of the current policy on achievement.

2. For content areas other than English (e.g., math, science), ELLs in this district may receive instruction that uses sheltered English immersion or Specially Designed Academic Instruction in English, which are intended to make the content accessible to ELLs; in this district, teachers of R-FEPs typically do not use these strategies. However, the district policies regarding sheltered English immersion and Specially Designed Academic Instruction in English are not as clear cut as those concerning ELD. Moreover, the focus of this study is on English language arts outcomes, and thus the primary service change of interest is the removal of ELD services.

3. Some of the studies reviewed refer to ELD and others to ESL. Gersten and Baker (2000) described the differences between ELD and ESL as variation in nomenclature across the country. To remain true to the authors of the original research cited here, I follow their use of the two terms. The district used for the empirical analyses of this study refers to these services as ELD.

4. One caveat, however, is that propensity score matching is most likely to yield estimates similar to a randomized experiment when (a) the full selection process of individuals into (in this case) ESL and non-ESL is known and modeled and (b) the matching is done locally (Cook, Shadish, & Wong, 2008). Callahan et al. (2009, 2010) matched students nationally (not locally), and the set of variables included in their matching does not account for all possible selection mechanisms; in particular, teacher selection of students on the basis of unobserved factors cannot be accounted for in their design.

5. Matsudaira’s (2009) district used either bilingual education services or ESL services, but Matsudaira did not estimate effects separately by the type of instructional services.

6. There are several key differences between Matsudaira’s (2009) study and the current study. First, Matsudaira looked at the initial year of ELD or bilingual service provision, while in the current study, I examine the last year of the bundle of services associated with the nonreclassified context. Second, Matsudaira’s conceptual focus was on estimating the effect of bilingual education, while the conceptual

### TABLE B3

**Robustness Check for Evidence of a Discontinuity in the Reduced-Form Equation of the Binding-Score Nonparametric Approach, Estimated at Various Points Away From the True Cut Score**

<table>
<thead>
<tr>
<th>Estimated at Cut Score of</th>
<th>Year of</th>
<th>Year after</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50 standard deviations below true cut score</td>
<td>−0.05 (0.07)</td>
<td>−0.04 (0.07)</td>
</tr>
<tr>
<td>0.25 standard deviations below true cut score</td>
<td>−0.09 (0.07)</td>
<td>−0.04 (0.08)</td>
</tr>
<tr>
<td>True cut score</td>
<td>−0.05 (0.05)</td>
<td>−0.16* (0.07)</td>
</tr>
<tr>
<td>0.25 standard deviations above true cut score</td>
<td>0.10 (0.10)</td>
<td>0.07 (0.17)</td>
</tr>
<tr>
<td>0.50 standard deviations above true cut score</td>
<td>0.06 (0.13)</td>
<td>0.06 (0.17)</td>
</tr>
</tbody>
</table>

*Note. Estimates are presented in pooled standard deviation units and can be interpreted as effect sizes. Robust standard errors appear in parentheses. The sample size for this high school level sample was 4,786. These nonparametric reduced-form regression discontinuity models use Imbens’s (2009) stata rdob ado file and include covariates for gender, special education status, race (Asian or Hispanic), lunch program status, and year fixed effects. Estimates at placebo cut scores use data from only their respective sides of the true cut score, as suggested by Imbens and Lemieux (2008). *5%.
focus or framing of the current study is on the district’s policy choice of the assessment threshold to provide a smooth transition from ELL to R-FEP. Third, with respect to the analytic method, the school district in Matsudaira’s analysis uses only one running variable to determine treatment (bilingual education) status (i.e., what Reardon and Robinson [in press] refer to as a single–rating score RD); therefore, he did not use a multiple–rating score RD, as I do in this study. Finally, although Matsudaira acknowledged that an IV would yield the treatment effect of interest, he instead opted to omit the IV and estimate an “intention to treat.”

7. Appendix B gives the estimated changes in the probability of reclassification for just attaining the final assessment threshold. That is, Table B1 gives the estimated coefficients on C from the first-stage equations (equation 1).

8. To check for manipulation, I used McCrary’s (2008) suggestion of exploring differences in sample density near the cut score. There was no evidence of this sort of manipulation. Following Robinson and Reardon’s (2009) suggestion for binding-score RD, I explored whether the tests contributing the binding score varied sharply at the cutoff. There was some evidence that CELDT listening/speaking tests were more often the binding score for students just above the threshold. In light of this, two strategies were used. First, CELDT listening/speaking and the overall CELDT (which partly reflects the listening/speaking score) were dropped from consideration as contributors to the binding score. This effectively introduced random noise into the model and biased estimates toward zero, as expected, though the signs of all the estimates remained intact. The second approach involved adding indicators of which test contributes the binding score into the estimation models. Doing so did not alter the estimated effects of reclassification. To further check for manipulation, I explored the smoothness of baseline (i.e., fall prior to the reclassification decision) grade point average when available (Grades 7–10) and found no evidence of manipulation. Therefore, there does not seem to be any strong evidence that manipulation occurred and led to biased effect estimates.

9. These observations all have 1 year of prior test scores and 1 year of follow-up test scores.

10. Note that reclassification policies did not change in this district during this time period.

11. Separate analyses were conducted for Asian and Hispanic students. The results were similar and statistically indistinguishable. Consequently, this article includes only the combined results for Asian and Hispanic students.

12. The lower mean CST-ELA scores at increasing grade levels may be the result of the accumulation of limited exposure to grade-level appropriate instruction in English language arts (cf. Callahan, 2005; Gándara et al., 2003; Olsen, 1997; Parrish et al., 2006). Such an interpretation highlights the complexity of reclassification, in that the objective should be to provide ELD services while still providing full access to the English language arts curriculum. Yet even this interpretation does not suggest that reclassification is necessarily “overdue.” Simply put, this pattern of decreasing mean CST-ELA scores is not a causal estimate, and we do not know what would happen to similar students in the ELD track compared with those in the non-ELD track. Therefore, these means are purely descriptive of the characteristics of students in the analytic samples at different grades, and I do not speculate about the causes of these trends.

13. In a separate set of OLS analyses, the binding-score terms were excluded, as such models may be a more typical or standard way of conditioning out prior achievement differences between ELLs and R-FEPs. These models yielded estimates almost identical to the OLS models included in this article.

14. The nonparametric-based estimates use local linear regression with a rectangular kernel and optimal bandwidth as determined by Guido Imbens’s (2009) rdrob Stata code and condition on all the same other covariates (e.g., gender, free and reduced-price lunch participation).

References


Author

JOSEPH P. ROBINSON, Ph.D., is an assistant professor of quantitative and evaluative research methodologies at the University of Illinois at Urbana-Champaign, Department of Educational Psychology, 210F Education Building, 1310 S. 6th Street, Champaign, IL 61820; jpr@illinois.edu. His research focuses on causal inference and quasi-experimental designs, policy analysis and program evaluation, and issues related to educational equity and access.

Manuscript received April 4, 2010
Revision received January 15, 2011
Accepted March 22, 2011